Cass means business

Cass Business School
City of London

Faculty of Actuarial Science and Statistics

The Quest for a Donor: Probability Based Methods Offer Help.

Petter.F. Mostad, Thore Egeland, Robert G. Cowell, Vidar Bosnes and Øivind Braaten.

Statistical Research Paper No. 26

November 2005

ISBN 1-901615-92-8

Cass Business School
106 Bunhill Row
London EC1Y 8TZ
T +44 (0)20 7040 8470
www.cass.city.ac.uk

The quest for a donor: Probability based methods offer help

Petter F. Mostad[1*†] ,Thore Egeland[2†], Robert G. Cowell[3], Vidar Bosnes[4] and Øivind Braaten[2]

[1]*Department of Biostatistics, University of Oslo, Box 1122 Blindern, 0317 Oslo, Norway*
[2]*Department of Medical Genetics, Ullevaal University Hospital, 0407 Oslo, Norway*
[3]*Faculty of Actuarial Science and Statistics, Cass Business School, London, UK*
[4]*Department of Immunology and Transfusion Medicine, Ullevaal University Hospital, 0407 Oslo, Norway*

[*]Correspondence to Petter F. Mostad, Department of Biostatistics,
University of Oslo, Box 1122 Blindern, 0317 Oslo Norway
Telephone: +47 2285 1399
Fax: +47 2285 1313
Email: p.f.mostad@medisin.uio.no

†P.F.M. and T.E. contributed equally to this study

Summary

When a patient in need of a stem cell transplant has no compatible donor within his or her closest family, and no matched unrelated donor can be found, a remaining option is to search within the patient's extended family. This situation often arises when the patient is of an ethnic minority, originating from a country that lacks a well-developed stem cell donor program, and has HLA haplotypes that are rare in his or her country of residence. Searching within the extended family may be time-consuming and expensive, and tools to calculate the probability of a match within groups of untested relatives would facilitate the search.

We present a general approach to calculating the probability of a match in a given relative, or group of relatives, based on the pedigree, and on knowledge of the genotypes of some of the individuals. The method extends previous approaches by allowing the pedigrees to be consanguineous and arbitrarily complex, with deviations from Hardy-Weinberg equilibrium. We show how this extension has a considerable effect on results, in particular for rare haplotypes. The methods are exemplified using freeware programs to solve a case of practical importance.

Keywords: Donor search; pedigree calculations; consanguinity

Introduction

Genetically matching stem cell donors can be found by searching within the patient's family, or by searching in bone marrow donor registries. The former approach is advantageous because relatives found to match the patient at tested HLA loci are much more likely to share the entire haplotypes with the patient than are matched unrelated donors, and this may improve the prognosis.[1] For patients with rare genotypes, searching within the family may be the only option. Consanguinity within the patient's pedigree, or sometimes occurrence of such events as two brothers marrying two sisters, may make finding a matching relative reasonably probable even when the patient's genotype is extremely rare. In fact, it might be this very consanguinity that has contributed to the occurrence of the rare genotype of the patient.

Rare haplotypes frequently complicate the search for donors to patients of ethnic minorities. Searching within the patient's extended family may be complex and expensive, as it may be scattered over several countries, and some family members may be difficult to reach and test. Thus it becomes important to optimize the search in terms of cost. An important part of such optimization is probability calculations. In particular, we need to (1) calculate the probability that another specified relative will match, i.e., share the relevant parts of the genotype of the patient, and (2) calculate the probability that at least one person in a specified group of relatives will match. The second calculation is important because a sensible search strategy typically involves testing relatives in groups, before re-evaluating the search plan.

Similar problems have been considered previously under simplifying assumptions, i.e., when the pedigrees have certain specified forms, when there is no inbreeding, and when the Hardy-Weinberg Equilibrium (HWE) holds. Methods, examples and a computer program called ExtFam have been described by Schipper et. al[1], see also Kollman[2] and Kaufman.[3]

We have found that methods and programs originally developed in forensic genetics can be applied to the present donor-matching problem, yielding solutions with greater generality and fewer assumptions.[4,5,6] This paper presents such methods and programs for doing the necessary probability calculations. We show how probabilities are changed considerably compared to those obtained with more simplistic methods, in particular when the haplotype frequencies are low.

Methods

Let *R* be a fixed pedigree, in which *n* persons $x_1,...,x_n$ have been tested, and let x_1 denote the patient. Let g(x) represent the genotype of person x, and let $g_1,...,g_n$ be the (known) genotypes of individuals $x_1,...,x_n$. We need to calculate the probability that a given relative x_{n+1} matches, i.e.

$$P\left(g(x_{n+1}) = g_{match} \mid g(x_1) = g_1,..., g(x_n) = g_n, R\right)$$

where g_{match} may represent g_1, or a genotype sufficiently close to g_1. It follows immediately from the definition of conditional probability that

$$= \frac{P\left(g(x_{n+1}) = g_{match} \mid g(x_1) = g_1,..., g(x_n) = g_n, R\right)}{P\left(g(x_{n+1}) = g_{match}, g(x_1) = g_1,..., g(x_n) = g_n \mid R\right)} \quad (1)$$

Thus we would like to compute quantities of the form

$$P\left(g(x_1) = g_1, g(x_2) = g_2,..., g(x_k) = g_k \mid R\right), \quad (2)$$

the probability of the occurrence of specified genotypes for *k* individuals within a given pedigree. Under the assumptions that there is no recombination within the relevant haplotypes in the pedigree, that the haplotypes involved can be deduced from the pedigree, that HWE holds, and that haplotype frequencies are available, (2) is possible to compute using standard probability calculations. Note that, in consanguineous pedigrees in particular, such computations can become quite complex, and disregarding the consanguinity leads to considerable errors. Although it is possible to derive formulas for each particular pedigree, it may be more practical to use a program implementing a Bayesian Network algorithm, which can handle any pedigree (see the Appendix).

The HWE assumption implies that if one of the haplotypes of a person is known, the probability distribution for the second haplotype remains unchanged. This is typically not the case when the parents of the person are on the average more related than two randomly selected persons from the population. Motivated by forensic applications, a practical way of dealing with such population substructure has been described by Balding and Nichols.[7] An important parameter is the "coancestry coefficient", or F_{ST}, which is the correlation between two haplotypes sampled from different individuals within a subpopulation. A positive correlation implies that the probability of observing a haplotype in a founder of a pedigree increases with every observation of this haplotype in the same or another founder. More precisely, if *s* founder haplotypes have been observed, and *r* of these are of type A, then the model implies that the probability for the next one to be of type A is

$$\frac{rF_{ST} + p_A(1-F_{ST})}{1+(s-1)F_{ST}} \tag{3}$$

For instance, the homozygote probability p_A^2 is replaced by $p_A(F_{ST} + (1-F_{ST})p_A)$. Observe that $F_{ST} = 0$ corresponds to HWE while positive values of F_{ST} increase the homozygote probability. The assumption of HWE has been implicit in previous treatments.[1,3]

The genotypes used in the computations could in principle be any set of allelic markers; the most relevant in this context are of course HLA haplotypes. Because of the extreme polymorphism of the HLA region, population frequencies can be very hard to estimate in the relevant populations. Consequently, it is of interest to discuss how to find probability bounds without using haplotype probabilities; in effect, finding the probability for a donor whose haplotypes are identical by descent within the pedigree to those of the patient. Standard probability calculations can be applied to this problem; see the Results section for an example.

Finally, let us return to the problem of calculating the probability of at least one match within a group of relatives. Let $x_{n+1},...,x_{n+k}$ be k relatives in the pedigree R in addition to the n considered above, and for $i = 1,...,k$ denote by m_i the event that $g(x_{n+i}) = g_{match}$, i.e., that this relative has the required genetic match. Then we would like to compute

$$P(m_1 \cup m_2 \cup ... \cup m_k \mid g(x_1) = g_1,..., g(x_n) = g_n, R). \tag{4}$$

Again, this can be computed in terms of quantities like (2), see the Appendix for details.

Results

We present three examples. First, the impact of deviation from HWE is studied. Then a practical case with a complex pedigree is studied assuming haplotype frequencies to be known. The last example expands on the second by assuming that the relevant haplotype frequency is unknown, but close to zero.

The effect of population substructure

For illustrative purposes, let us consider the case where a mother is considered as a donor for a homozygote AA child. When HWE is assumed, the probability for a match is p_A, the population frequency of A. When F_{ST} is nonzero, it follows from (1) and (3) that

$$P(\text{mother matches}) = \frac{2F_{ST} + p_A(1-F_{ST})}{1+F_{ST}}, \tag{5}$$

and the ratio of (5) to p_A is

$$\frac{2}{1+F_{ST}}\frac{F_{ST}}{p_A} + \frac{1-F_{ST}}{1+F_{ST}}.$$

Although F_{ST} is difficult to estimate and varies considerably among populations, a typical and conservative value could be 0.01.[8] When p_A is much larger than F_{ST} the ratio above is close to 1, and the adjustment does not matter. But when p_A is approximately equal in size to F_{ST}, as it may very well be for many HLA haplotypes, the ratio is around 3, indicating the necessity of estimating and using correct F_{ST} values in calculations.

Sample calculations in a consanguineous pedigree

Consider the family extending over four generations shown in Figure 1. The parents of the homozygote AA patient, IV1 and IV2, are cousins and so are the parents of the father, III3 and III4. Note that some persons in the pedigree are present only to define family relations, they are not available as potential donors. The patient's immediate family has been tested without finding a match.

We consider initially the patient's four grandparents as possible donors. Table 1 compares match probabilities for III1-III4 at different values of p_A. First, one calculates the match probability for a grandparent disregarding consanguinity (i.e., generations I and II are removed from the pedigree). Then, this is compared to the correct pedigree using F_{ST}=0, and using F_{ST}=0.01, for each of the four grandparents. The results were obtained using the

Familias program (see the Appendix). We see that disregarding the known ancestry of the grandparents can affect match probabilities substantially in both directions. Disregarding Hardy-Weinberg disequilibrium tends to lead to underestimation.

As an illustration, we also compute the probability that at least one of the grandparents, or at least one of the uncles, will match. Table 2 shows this probability for different values of p_A and under different hypotheses. See the Appendix on how the values have been obtained.

A lower bound on the probability of finding a matching donor in a consanguineous pedigree

Knowing about consanguinity within the pedigree may either increase or decrease the match probability for individual potential donors, as we saw above. However, persons who are inbred within the pedigree (i.e., whose two haplotypes may have been inherited from a single haplotype in the pedigree) generally increase their match probabilities to a homozygote patient substantially, in particular when p_A is low. In the example of Figure 1, consider the potential donors IV3-IV6. If the patient inherited one or both of her haplotypes A from a haplotype of I1 or I2, then any of IV3 to IV6 may have inherited two copies of A to become homozygous AA. Thus these persons have a match probability above a certain lower bound no matter how small p_A is.

In fact, the lower bound for the probability of at least one of IV3-IV6 being a matching donor, is

$$\frac{1}{9}\left[1-\left(\frac{3}{4}\right)^4\right] = 0.076. \qquad (6)$$

Table 2 shows matching probabilities at different p_A's, and under different hypotheses. The effect of considering consanguinity, and the presence of a lower bound, is very clear. To obtain the number given in (6), we first compute the probability that both III3 and III4 carry the A haplotype. As p_A may be arbitrarily low, we may disregard the possibility that A occurs more than once among the pedigree founders. Thus Figure 2 illustrates the ways A may have spread in the pedigree. A counting argument shows that the probability of both III3 and III4

carrying a copy of A is 1/9; see the Appendix for details. Simple Mendelian computations then give the result in (6).

Discussion

The examples above show that some care must be taken in computations, in order to get reliable numbers for match probabilities within the extended family. We have focused on including in the computations parts of the pedigree indicating consanguinity, and using an adjustment for population substructure. Below, we discuss some additional possible extensions of computational methods.

We have assumed that haplotypes have been observed, and have used these as data. However, the haplotypes are not directly observed, only the alleles at each of the usually three markers used in the HLA region. As the markers are very polymorphic and we generally start with data coming from several generations, it is usually possible to deduce the phase of the markers (assuming no recombinations have taken place). But in some examples, this may be impossible (e.g., if the mother, father, and child all have the same two markers in a locus). It is not too hard to extend computational methods to such cases: We just need to sum probabilities over the different possible haplotypes compatible with the observed data.

Even though the recombination probabilities between HLA loci are small,[9] a crossover will occur in a significant proportion of the pedigrees of extended families we consider. Such crossovers may not always be possible to deduce from the allelic data we observe. However, even undetected crossovers may have a decisive influence on the probability of finding a match in untested parts of the pedigree. A possible solution is to include the possibility of crossovers in the inheritance model. This is not too difficult in principle, but would require some reprogramming of existing tools. As the probability of undetected crossovers is low, the effect on the final result is likely to be limited.

Our methods, in the way we have described them, do not separate between haplotypes identical to those of the patient, and haplotypes merely identical in the three loci, even if the difference between these types can be one of the motivations for searching for a donor in the extended pedigree. It is however possible to treat these differently: We are then faced with

computing the probability of different kinds of matches, some considered better than others. Indeed, in cases where the probability of finding a match in all three systems is very low, one might want to search for a one-mismatch donor. The computations and methods for doing this would be more or less unchanged: One would just widen the definition of genetic match.

Throughout we have assumed fixed values for parameters, i.e., haplotype frequencies and F_{ST}. These fixed values may be replaced by probability distributions. The resulting answers would then include parameter uncertainty. This might be a relevant way to handle cases where population frequencies are very difficult to estimate. In cases where analytical formulae are available the extension is easy, whereas more general cases might require an extended Bayesian network to be programmed, or the application of simulation methods.

We have shown in this paper that computational models with fairly high complexity are necessary to obtain reliable numbers for donor match probabilities in some cases. In practical cases, it is then central to have available software tools making such computations feasible for the practitioner. The freely available program Familias[10] is one alternative, and a webpage with help for its use for the donor match application is available.[11] However, any tool implementing Bayesian Network computations could be used. Such general tools might supply answers to some questions more efficiently, but they might also require a more experienced user, in particular in implementing Hardy-Weinberg disequilibrium. Unfortunately, no program we know of is directly tailored for the applications in this paper, and making such an adaptation would be desirable.

Acknowledgements

The three first authors were supported by a Research Interchange Grant from the Leverhulme trust.

References

1. Schipper RF, D'Amaro J, Oudshoorn M. The probability of finding a suitable related donor for bone marrow transplantation in extended families. *Blood* 1996; **87:** 800-804.

2. Kollman C. Probability calculations for a matched donor in the extended family. *Blood* 1996; **87:** 5391-5392.
3. Kaufman R. A generalized HLA prediction model for related donor matches. *Bone Marrow Transplant* 1996; **17**: 1013-1020.
4. Egeland T, Mostad PF, Mevåg B, Stenersen M. Beyond traditional paternity and identification cases: Selecting the most probable pedigree. *Forensic Sci Int* 2000; **110**: 47-59.
5. Egeland T, Mostad PF. Statistical genetics and genetical statistics: a forensic perspective. *Scand J Statist* 2002; **29**: 297-307.
6. Dawid AP, Mortera J, Pascali VL, van Boxel D. Probabilistic expert systems for forensic inference from genetic markers. *Scand J Statist* 2002; **29**: 577-595.
7. Balding D, Nichols RA. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 1995; **96**: 3-12.
8. Evett IW, Weir BS. *Interpreting DNA Evidence.* Sinauer, Sunderland, MA, 1998.
9. Cullen M, Perfetto SP, Klitz W *et al*. High-Resolution Patterns of Meiotic Recombinations across the Human Major Histocompatibility Complex. *Am J Hum Genet* 2002; **71**: 759-776.
10. http://www.nr.no/familias
11. http://www.math.chalmers.se/~mostad/familias/donors
12. Feller W. *Probability theory and its applications.* Wiley, NY, 1959.
13. Cowell RG, Dawid AP, Lauritzen SL, Spiegelhalter DJ. *Probabilistic Networks and Expert Systems*. Springer-Verlag, Berlin-Heidelberg-New York, 1999.
14. http://www.cs.Berkeley.edu/~murphyk/Bayes/bnintro.html
15. Lauritzen SL, Sheehan NA. Graphical Models for Genetic Analyses. *Stat Sci* 2003; **18**: 489-514.

Appendix

In this Appendix, we provide some details on how to obtain the results in Tables 1 and 2 using various computer programs. We also add some detail to the computations leading to the result in (6).

Using the familias program

Familias[4,5] was developed primarily to investigate complex paternity cases, and other cases with competing pedigree hypotheses. However, it can also be used to compute quantities like (2), with a given value for F_{ST}. All the values in Table 1 were obtained according to Equation (1) as a quotient of two values computed by Familias.

It is unfortunately not as easy to obtain values like those in Table 2. A general approach to computing the probability of unions of events, as in Equation (4), is to use the Inclusion-Exclusion principle.[12] Let

$$\begin{aligned} S_1 &= \sum_{i_1} P(m_{i_1} \mid \bullet), \\ S_2 &= \sum_{i_1<i_2} P(m_{i_1}, m_{i_2} \mid \bullet) \\ &\vdots \\ S_k &= \sum_{i_1<i_2<\ldots<i_k} P(m_{i_1}, m_{i_2}, \ldots, m_{i_k} \mid \bullet) \end{aligned}$$

where "$\bullet$" is short for the available genotype information $g(x_1) = g_1, \ldots, g(x_n) = g_n$ and pedigree structure R. Then

$$P(m_1 \cup m_2 \cup \ldots \cup m_k \mid \bullet) = S_1 - S_2 + \ldots + (-1)^{k+1} S_k . \qquad (7)$$

The $2^k - 1$ terms in this result can easily become too many for manual computation, although useful bounds could be obtained by truncating to the first few terms. One simplification occurs if individuals 1,...,s are independent of s+1,...,k, for then

$$P(m_1 \cup m_2 \cup \ldots \cup m_k \mid \bullet) = P(m_1 \cup m_2 \cup \ldots \cup m_s \mid \bullet) - P(m_{s+1} \cup m_{s+2} \cup \ldots \cup m_k \mid \bullet) - P(m_1 \cup m_2 \cup \ldots \cup m_s \mid \bullet) P(m_{s+1} \cup m_{s+2} \cup \ldots \cup m_k \mid \bullet) \quad .$$

In many cases there might be other simplifying circumstances: For example, to compute the results on the left side of Table 2, we can use that at most one parent of each of IV1 and IV2 can be a match. Thus Equation (7) reduces to the sum of the match probabilities for each of the grandparents III1-III4, minus the probabilities for the 4 possible double matches: III1 and III3; III1 and III4; III2 and III3, and III2 and III4. The right part of table 3 can also be computed similarly, although all 15 terms of the original formula (7) must now be included. An alternative is to first compute the probability for the five possible relevant genotypes for III3 and III4: (AA, AB), (AB, AB), (AX, AB), (AB, AA), and (AB, AX), where X is any haplotype different from A and B. In the first and third of these cases, the probability for a match among IV3-IV6 is $1-\left(\frac{1}{2}\right)^4$, while in the second and fourth case, it is $1-\left(\frac{3}{4}\right)^4$. A webpage[11] is available for further details about applying Familias to the problems of this paper.

Using Bayesian Network (BN) software

Bayesian networks[13] are a type of directed graphical model that captures conditional independences between random variables in multivariate stochastic models. A web site with excellent tutorial material on Bayesian networks with a link to a comprehensive list of Bayesian network software (for example the commercial package HUGIN and the freeware GENIE) that could be used to perform the calculations described in this paper, is maintained by Kevin Murphy.[14] In using BN software for the problems tackled in this paper, there is some freedom in the choice of random variables that can be used to model the pedigree; here for simplicity we shall use the genotype representation, in which only the genotypes of the people in the pedigree are represented by random variables in the Bayesian network (other representations can be much more efficient computationally).[15] Figure 3a shows the Bayesian network for a first simple example, assuming HWE. The nodes represent the genotypes of the grandmother, mother and child. Associated with the grandmother node is the prior distribution for the genotype that the grandmother could have. Associated with the mother node is a conditional probability for the mother's genotype given the grandmother's genotype; a distribution with identical values is specified for the child node. After setting the network up in a BN, evidence consisting of the child's genotype is entered on the child node and

distributed to the other nodes of the network by a process called 'propagation'. After this (usually automatic) propagation process the marginal posterior distributions of the mother and grandmother genotypes are displayed, and from these can be read off the probabilities for the mother or grandmother to individually match the child.

Figure 3b shows a Bayesian network to calculate these posterior probabilities taking into account deviations from Hardy-Weinberg equilibrium. For this calculation, we need to introduce the child's father and maternal grandfather as extra nodes in the network. In order to capture the correlations between haplotypes, at the top of the network is another random variable whose states consist of the set of all combinations of the genotypes of the father, grandmother and grandfather; there is an associated probability table that models the deviations from HWE as described by Balding and Nichols.[7] The conditional probabilities associated with the father, grandmother and grandfather nodes will take value of either 0 or 1 to reflect logical constraints, for example, that the genotype of the father in the f node must match that in the joint node. The conditional probabilities specified for the mother and child nodes model Mendel's law of gene inheritance. Using this network the probability that each of the mother or grandmother individually matches can be found in exactly the same way as for the simpler network of Figure 3a. The complexity of this approach increases substantially with the number of observed haplotypes in the marker, and the number of founders in the pedigree. For example, to do the calculations in Table 1 in BN software would require a node to represent the joint genotype of five founders, the probability table to be specified has 100,000 entries, and the conditional probability table for each of the five founders given this joint variable would have one million entries. Clearly it is not feasible to specify this manually using BN software – some programming would be required. Hence such corrections are best evaluated using Familias. (In contrast, assuming HWE, the complexity is much smaller, with the largest conditional probability table having 1000 entries, a number that can be reduced to 250 using an alternative Bayesian network representation of the pedigree.)

Using BN software to find probabilities like (4) is a two-step process. After setting up the pedigree (either a simple network assuming HWE or a more complex network that incorporates HWE deviations) one enters the haplotype data on the patient, her three siblings and her parents and propagates. BN software returns the probability of this data, $P(\bullet)$ as a so-called normalization constant. Next one enters as additional (likelihood) evidence the restriction that each of the people in the group is NOT a match; then the normalization

constant is $P(\overline{m_1}, \overline{m_2}, ..., \overline{m_k}, \bullet)$, the joint probability of the observed genotypes of the six haplotyped people and also the restriction that nobody in the group matches the patient. Finding the probability of at least one match in the group is then an application of basic probability theory:

$$P(m_1 \cup m_2 \cup ... \cup m_k \mid \bullet) = 1 - P(\overline{m_1}, \overline{m_2}, ..., \overline{m_k} \mid \bullet) = 1 - \frac{P(\overline{m_1}, \overline{m_2}, ..., \overline{m_k}, \bullet)}{P(\bullet)}.$$

That is, the ratio of the second normalization constant to the first, subtracted from 1, will give the probability that there is at least one match among the people in the group of interest.

Computing lower bounds for match probabilities

We conclude with some details on the computations leading to the number given in (6). In Figure 2, the pedigrees (ii) through (v) have companions where A appears in I2 instead of I1, giving a total of 9 possible ways A can spread in the pedigree. What is the probability of observing the data in each case? Note that for all persons except the founder it is known whether A is the paternal or maternal haplotype; assume first that it is the maternal haplotype in the founder. The probability of observing these data is p_A times $1 - p_A$ to the power of the number of other founding haplotypes, which is 9, times ½ to the power of the number of times a particular choice of paternal or maternal haplotype must have been made in order to give rise to the given data. In example (i) there are 4 such segregation events, so the probability of the data is $p_A(1-p_A)^9(1/2)^4$. The case that the founding A haplotype is the paternal haplotype of the founder has exactly the same probability, so the probability of observing the data indicated in Figure 2(i) is $2p_A(1-p_A)^9(1/2)^4$. We have in fact also observed that the patient's mother has a B haplotype and that the patient's father has a C haplotype. The mother's maternal haplotype is a founder haplotype, and by summing over the possibilities for how the father's paternal haplotype may have been inherited, we get that the probability for observing the actual data is

$$2p_A p_B p_C (1-p_A)^7 \left(\frac{1}{2}\right)^4.$$

Exactly the same argument may be repeated for the other pedigrees in Figure 2, so that the total probability of observing any of them is

$$2p_A p_B p_C (1-p_A)^7 \left[\frac{1}{2^4} + 2\frac{1}{2^6} + 2\frac{1}{2^7} + 2\frac{1}{2^7} + 2\frac{1}{2^7} \right] = 2p_A p_B p_C (1-p_A)^7 \frac{9}{64}. \qquad (8)$$

Of the pedigrees in Figure 2, only in type (v) do the parents III3 and III4 both carry haplotype A, so this happens with probability

$$2p_A p_B p_C (1-p_A)^7 2\frac{1}{2^7} = 2p_A p_B p_C (1-p_A)^7 \frac{1}{64}. \qquad (9)$$

The conditional probability that these parents both carry A, given the observed data, is then the ratio of the numbers computed in (8) and (9), thus 1/9.

Tables

Table 1. Match probabilities for the grandparents III1-III4 from the pedigree in Figure 1 under various models.

p_A	Grandparent								
	Any	III1		III2		III3		III4	
	Model								
	Simplified[a]	HWE	F_{ST}=0.01	HWE	F_{ST}=0.01	HWE	F_{ST}=0.01	HWE	F_{ST}=0.01
0.1[b]	0.05[b]	0.00046	0.093	0.18	1.8	0.15	1.6	0.033	0.41
0.3	0.15	0.0041	0.12	0.52	2.1	0.46	1.8	0.10	0.48
1.0	0.5	0.043	0.22	1.7	3.1	1.4	2.7	0.34	0.74
3.0	1.5	0.33	0.64	4.4	5.5	3.8	4.8	1.1	1.5
10	5.0	2.4	2.9	10.7	11.3	9.5	10.1	3.8	4.3

[a]In the simplified model, consanguinity is ignored (generations I and II are removed from the pedigree), and HWE is assumed ($F_{ST} = 0$).

[b]Frequencies and match probabilities are shown as percentages.

Table 2. Match probabilities for at least one of the grandparents III1-III4, or at least one of the uncles IV3-IV6, under various models.

p_A	Match in at least one grandparent			Match in at least one of four uncles		
	Model			Model		
	Simplified[a]	HWE	F_{ST} = 0.01	Simplified	HWE	F_{ST} = 0.01
0.1[b]	0.20[b]	0.29	3.1	0.068	7.7	8.7
0.3	0.60	0.85	3.6	0.21	7.9	8.9
1.0	2.0	2.8	5.4	0.69	8.6	9.5
3.0	5.9	7.6	9.9	2.1	10	11
10	19	21	23	7.1	16	16

[a]In the simplified model, consanguinity is ignored (generations I and II are removed from the pedigree), and HWE is assumed ($F_{ST} = 0$).

[b]Frequencies and match probabilities are shown as percentages.

Figure 1: The family extended over four generations, V4 represents the patient seeking a donor, with siblings V1-V3 and parents IV1 and IV2 having already been typed and found not to match.

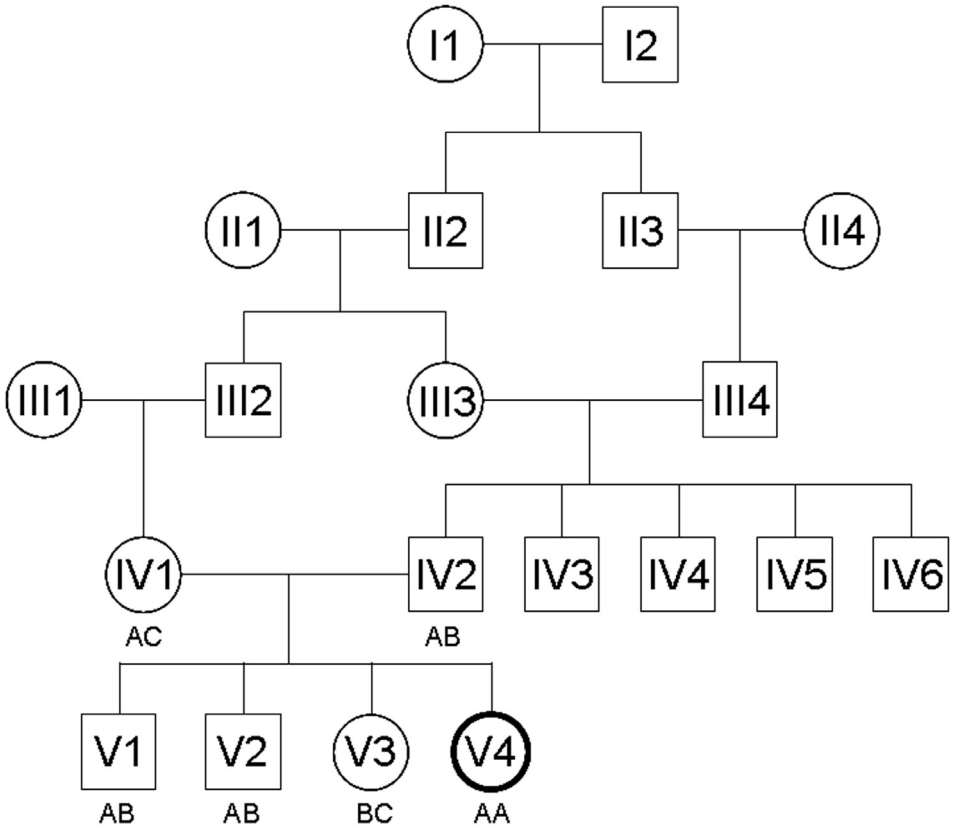


Figure 2: Illustrations of the ways in which the allele A could have been passed down from a single founder in the pedigree.

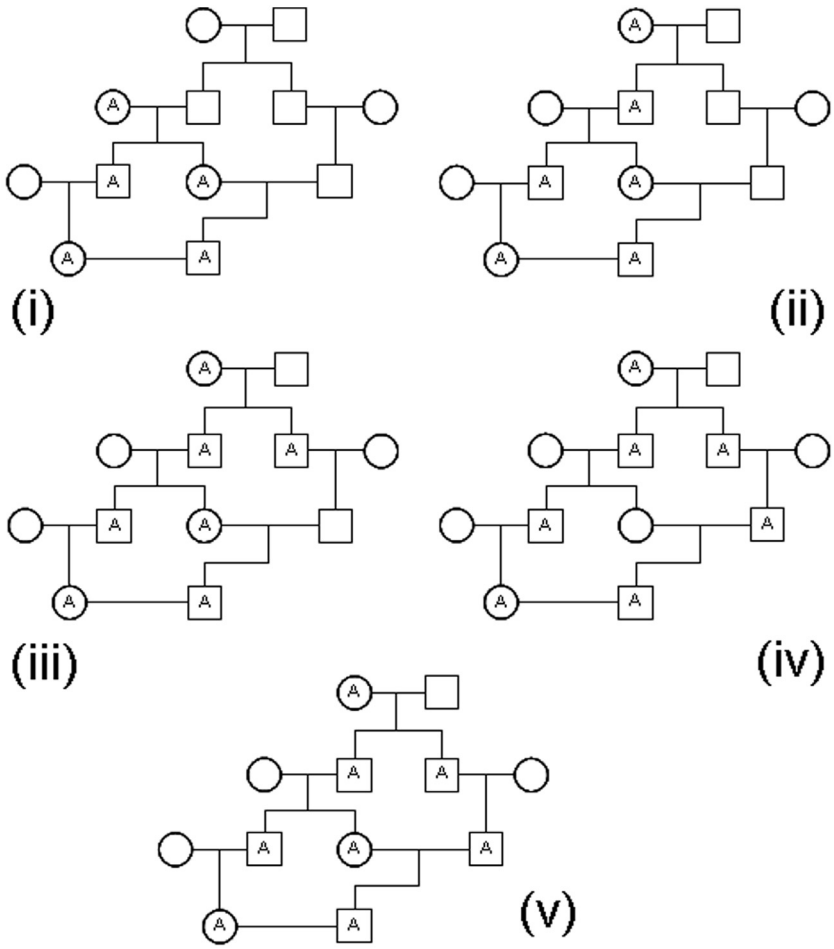
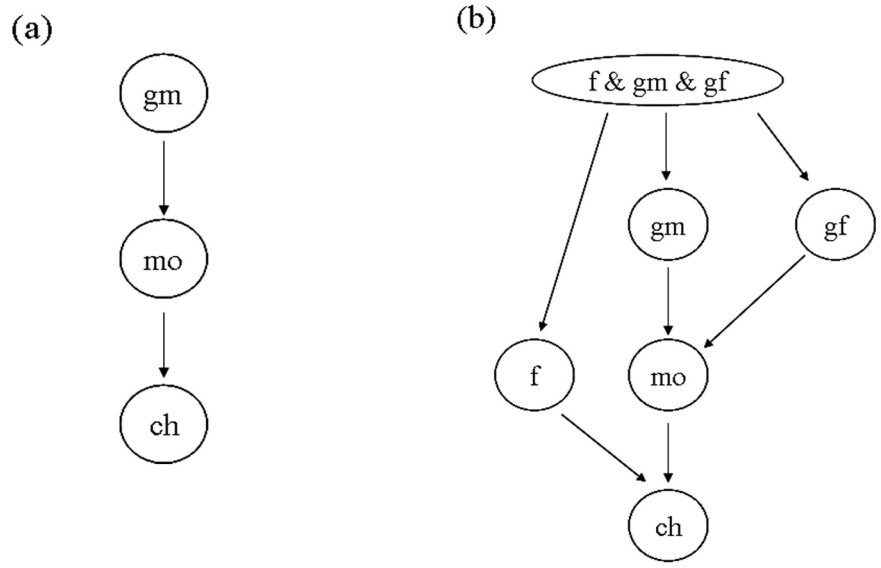


Figure 3: (a) Bayesian network for a simple example, assuming Hardy-Weinberg equilibrium; (b) Bayesian network to take into account deviations from Hardy-Weinberg equilibrium.



Cass Business School
City of London

FACULTY OF ACTUARIAL SCIENCE AND STATISTICS

Actuarial Research Papers since 2001

135. Renshaw A. E. and Haberman S. On the Forecasting of Mortality Reduction Factors. February 2001.
ISBN 1 901615 56 1

136. Haberman S., Butt Z. & Rickayzen B. D. Multiple State Models, Simulation and Insurer Insolvency. February 2001. 27 pages.
ISBN 1 901615 57 X

137. Khorasanee M.Z. A Cash-Flow Approach to Pension Funding. September 2001. 34 pages.
ISBN 1 901615 58 8

138. England P.D. Addendum to "Analytic and Bootstrap Estimates of Prediction Errors in Claims Reserving". November 2001. 17 pages.
ISBN 1 901615 59 6

139. Verrall R.J. A Bayesian Generalised Linear Model for the Bornhuetter-Ferguson Method of Claims Reserving. November 2001. 10 pages.
ISBN 1 901615 62 6

140. Renshaw A.E. and Haberman. S. Lee-Carter Mortality Forecasting, a Parallel GLM Approach, England and Wales Mortality Projections. January 2002. 38 pages.
ISBN 1 901615 63 4

141. Ballotta L. and Haberman S. Valuation of Guaranteed Annuity Conversion Options. January 2002. 25 pages.
ISBN 1 901615 64 2

142. Butt Z. and Haberman S. Application of Frailty-Based Mortality Models to Insurance Data. April 2002. 65 pages.
ISBN 1 901615 65 0

143. Gerrard R.J. and Glass C.A. Optimal Premium Pricing in Motor Insurance: A Discrete Approximation. (Will be available 2003).

144. Mayhew, L. The Neighbourhood Health Economy. A systematic approach to the examination of health and social risks at neighbourhood level. December 2002. 43 pages.
ISBN 1 901615 66 9

145. Ballotta L. and Haberman S. The Fair Valuation Problem of Guaranteed Annuity Options: The Stochastic Mortality Environment Case. January 2003. 25 pages.
ISBN 1 901615 67 7

146. Haberman S., Ballotta L. and Wang N. Modelling and Valuation of Guarantees in With-Profit and Unitised With-Profit Life Insurance Contracts. February 2003. 26 pages.
ISBN 1 901615 68 5

147. Ignatov Z.G., Kaishev V.K and Krachunov R.S. Optimal Retention Levels, Given the Joint Survival of Cedent and Reinsurer. March 2003. 36 pages.
ISBN 1 901615 69 3

148. Owadally M.I. Efficient Asset Valuation Methods for Pension Plans. March 2003. 20 pages.
ISBN 1 901615 70 7

149. Owadally M.I. Pension Funding and the Actuarial Assumption Concerning Investment Returns. March 2003. 32 pages.

ISBN 1 901615 71 5

150. Dimitrova D, Ignatov Z. and Kaishev V. Finite time Ruin Probabilities for Continuous Claims Severities. Will be available in August 2004.

151. Iyer S. Application of Stochastic Methods in the Valuation of Social Security Pension Schemes. August 2004. 40 pages.

ISBN 1 901615 72 3

152. Ballotta L., Haberman S. and Wang N. Guarantees in with-profit and Unitized with profit Life Insurance Contracts; Fair Valuation Problem in Presence of the Default Option[1]. October 2003. 28 pages.

ISBN 1-901615-73-1

153. Renshaw A. and Haberman. S. Lee-Carter Mortality Forecasting Incorporating Bivariate Time Series. December 2003. 33 pages.

ISBN 1-901615-75-8

154. Cowell R.G., Khuen Y.Y. and Verrall R.J. Modelling Operational Risk with Bayesian Networks. March 2004. 37 pages.

ISBN 1-901615-76-6

155. Gerrard R.G., Haberman S., Hojgaard B. and Vigna E. The Income Drawdown Option: Quadratic Loss. March 2004. 31 pages.

ISBN 1-901615-77-4

156. Karlsson, M., Mayhew L., Plumb R, and Rickayzen B.D**.** An International Comparison of Long-Term Care Arrangements. An Investigation into the Equity, Efficiency and sustainability of the Long-Term Care Systems in Germany, Japan, Sweden, the United Kingdom and the United States. April 2004. 131 pages.

ISBN 1 901615 78 2

157. Ballotta Laura. Alternative Framework for the Fair Valuation of Participating Life Insurance Contracts. June 2004. 33 pages.

ISBN 1-901615-79-0

158. Wang Nan. An Asset Allocation Strategy for a Risk Reserve considering both Risk and Profit. July 2004. 13 pages.

ISBN 1 901615-80-4

159. Spreeuw Jaap. Upper and Lower Bounds of Present Value Distributions of Life Insurance Contracts with Disability Related Benefits. December 2004. 35 pages.

ISBN 1 901615-83-9

160. Renshaw A.E. and Haberman S. Mortality Reduction Factors Incorporating Cohort Effects. January 2005. 33 pages.

ISBN 1 90161584 7

161. Gerrard R.J. Haberman A and Vigna E. The Management of De-Cumulation Risks in a Defined Contribution Environment. February 2005. 35 pages.

ISBN 1 901615 85 5.

162. Ballotta L, Esposito G. and Haberman S. The IASB Insurance Project for Life Insurance Contracts: Impart on Reserving Methods and Solvency Requirements. May 2005. 26 pages.

ISBN 1-901615 86 3.

163. Emms P. and Haberman S. Asymptotic and Numerical Analysis of the Optimal Investment Strategy for an Insurer. September 2005. 42 pages.

ISBN 1-901615-88-X

164. Kaishev V.K., Dimitrova D.S. and Haberman S Modelling the Joint Distribution of Competing Risks Survival Times using Copula Functions. October 2005. 26 pages.

ISBN 1-901615-89-8

165. Kaishev V.K. and Dimitrova D.S. Excess of Loss Reinsurance Under Joint Survival Optimality. November 2005. 18 pages.

ISBN1-901615-90-1

166. Biffis E. and Denuit M. Lee-Carter Goes Risk-Neutral. An Application to the Italian Annuity Market. November 2005. 22 pages.

ISBN 1-901615-91-X

Statistical Research Papers

1. Sebastiani P. Some Results on the Derivatives of Matrix Functions. December 1995. 17 Pages.

ISBN 1 874 770 83 2

2. Dawid A.P. and Sebastiani P. Coherent Criteria for Optimal Experimental Design. March 1996. 35 Pages.

ISBN 1 874 770 86 7

3. Sebastiani P. and Wynn H.P. Maximum Entropy Sampling and Optimal Bayesian Experimental Design. March 1996. 22 Pages.

ISBN 1 874 770 87 5

4. Sebastiani P. and Settimi R. A Note on D-optimal Designs for a Logistic Regression Model. May 1996. 12 Pages.

ISBN 1 874 770 92 1

5. Sebastiani P. and Settimi R. First-order Optimal Designs for Non Linear Models. August 1996. 28 Pages.

ISBN 1 874 770 95 6

6. Newby M. A Business Process Approach to Maintenance: Measurement, Decision and Control. September 1996. 12 Pages.

ISBN 1 874 770 96 4

7. Newby M. Moments and Generating Functions for the Absorption Distribution and its Negative Binomial Analogue. September 1996. 16 Pages.

ISBN 1 874 770 97 2

8. Cowell R.G. Mixture Reduction via Predictive Scores. November 1996. 17 Pages.

ISBN 1 874 770 98 0

9. Sebastiani P. and Ramoni M. Robust Parameter Learning in Bayesian Networks with Missing Data. March 1997. 9 Pages.

ISBN 1 901615 00 6

10. Newby M.J. and Coolen F.P.A. Guidelines for Corrective Replacement Based on Low Stochastic Structure Assumptions. March 1997. 9 Pages.

ISBN 1 901615 01 4.

11. Newby M.J. Approximations for the Absorption Distribution and its Negative Binomial Analogue. March 1997. 6 Pages.

ISBN 1 901615 02 2

12. Ramoni M. and Sebastiani P. The Use of Exogenous Knowledge to Learn Bayesian Networks from Incomplete Databases. June 1997. 11 Pages.

ISBN 1 901615 10 3

13. Ramoni M. and Sebastiani P. Learning Bayesian Networks from Incomplete Databases. June 1997. 14 Pages.

ISBN 1 901615 11 1

14. Sebastiani P. and Wynn H.P. Risk Based Optimal Designs. June 1997. 10 Pages.
ISBN 1 901615 13 8

15. Cowell R. Sampling without Replacement in Junction Trees. June 1997. 10 Pages.
ISBN 1 901615 14 6

16. Dagg R.A. and Newby M.J. Optimal Overhaul Intervals with Imperfect Inspection and Repair. July 1997. 11 Pages. ISBN 1 901615 15 4

17. Sebastiani P. and Wynn H.P. Bayesian Experimental Design and Shannon Information. October 1997. 11 Pages. ISBN 1 901615 17 0

18. Wolstenholme L.C. A Characterisation of Phase Type Distributions. November 1997. 11 Pages. ISBN 1 901615 18 9

19. Wolstenholme L.C. A Comparison of Models for Probability of Detection (POD) Curves. December 1997. 23 Pages. ISBN 1 901615 21 9

20. Cowell R.G. Parameter Learning from Incomplete Data Using Maximum Entropy I: Principles. February 1999. 19 Pages. ISBN 1 901615 37 5

21. Cowell R.G. Parameter Learning from Incomplete Data Using Maximum Entropy II: Application to Bayesian Networks. November 1999. 12 Pages ISBN 1 901615 40 5

22. Cowell R.G. FINEX : Forensic Identification by Network Expert Systems. March 2001. 10 pages.
ISBN 1 901615 60X

23. Cowell R.G. When Learning Bayesian Networks from Data, using Conditional Independence Tests is Equivalant to a Scoring Metric. March 2001. 11 pages.
ISBN 1 901615 61 8

24. Kaishev, V.K., Dimitrova, D.S., Haberman S., and Verrall R.J. Autumatic, Computer Aided Geometric Design of Free-Knot, Regression Splines. August 2004. 37 pages.
ISBN 1-901615-81-2

25. Cowell R.G., Lauritzen S.L., and Mortera, J. Identification and Separation of DNA Mixtures Using Peak Area Information. December 2004. 39 pages.
ISBN 1-901615-82-0

26. Mostad P.F., Egeland T., Cowell R.G., Bosnes V. and Braaten Ø. The Quest for a Doner : Probability Based Methods Offer Help. November 2005. 19 Pages.
ISBN 1-90161592-8

Faculty of Actuarial Science and Statistics

Actuarial Research Club

The support of the corporate members

CGNU Assurance
English Matthews Brockman
Government Actuary's Department
Watson Wyatt Partners

is gratefully acknowledged.

ISBN 1-901615-92-8